NASA-TM-112653

RND-88-002

# MASS STORAGE REQUIREMENTS IN A COMPUTATIONAL FLUID DYNAMICS ENVIRONMENT

Frank S. Preston

Numerical Aerodynamics Simulation Systems Division NASA Ames Research Center, Moffett Field, California 94035

## ABSTRACT

**Computational Fluid Dynamics (CFD) research and applications produce massive solution files, as do other similar scientific analysis which employ large three dimensional grids and iterative solutions. This imposes mass storage requirements that exceed the capability of present practical storage technology and economic limits. A single solution will produce 40 MegaBytes of output. A time dependent solution may need 1,000 such solution steps or 40 GigaBytes. One user could generate more than this in a week. For a project lasting a year, this user would like to save in excess of 500 GigaBytes. Even if only 10% of the Numerical Aerodynamics Simulation (NAS) 600+ users were so prolific, this would total 30 TeraBytes. With today's technology, the practical solution for Mass Storage depends on developing capability and procedures to make the most effective use of a constrained resource. Workload and system models are being employed on the NAS program to aid in the design for a large increase in Mass Storage capability.**

## INTRODUCTION

Supercomputer environments employ storage at various levels in a hierarchical fashion. For this paper these will be differentiated by different names. The disk storage of the supercomputer which works with

the main memory will be referred to as the working store. The longer term storage generally consists of disks and tapes. Although these work together, they serve different functions, have different access times, and employ different technologies. The term Mass Storage will refer to the on-line disks, and the term Archive for the tapes. This paper concentrates on the mass storage devoted to the main, permanent storage part of the hierarchy.

The Numerical Aerodynamics Simulation (NAS) System serves a wide variety of scientific users at several NASA centers, other governmental activities, many Universities, and a number of commercial facilities. About one hundred users are located at Ames Research Center and about 600 are at remote locations served by a variety of directly connected communication lines. The system is known as the NAS Processing System Network (NPSN).

Supercomputer users and their problems impose super-large demands on the systems. These reflect in enormous requirements for storage. Although NAS users are working on a wide spectrum of scientific applications, the principal focus is on Computational Fluid Dynamics (CFD). This paper will use CFD as an example of a class of applications that push requirements beyond present economic and space constraints. Users' demands would now require an approach which looks like infinity to the system designer.

The NPSN is structured [1] to employ two supercomputers with one being at the leading edge of technology and the other from a previous more mature generation. Two Cray-2's are installed at present. (One is an interim machine which will be replaced later with a next generation computer.) In addition, specialized computers are supported such as a Connection Machine plus a number of mini-supercomputers. The Cray-2's have a main memory of 256 million 64 bit words and users solutions can easily exceed 50 million words. The working disk storage is about 40 GigaBytes. To support the supercomputers, files are moved over high speed trunks to a Mass Storage Subsystem (MSS) consisting of high speed magnetic disks. This is now 120 GigaBytes and will be increased to 240 GB or more. For an Archive, tape cartridges are used which provide access latency of less than a few minutes. Tapes are largely used for input files to start projects and outgoing final results for remote user archives, etc. A block diagram of the NAS system is shown in Figure 1.

The NAS system is being increased (1988-89) in processor performance by a factor of about five. Experience at supercomputer centers has shown

that this generally results in at least a corresponding increase in network traffic and mass storage needs. The designers are now increasing the access speed. Total storage will be increase as higher density drives become available. Even with all the possible increased mass storage capability, the amount possible is less than real and immediate requirements. Therefore, it is necessary to develop alternative strategies to handle the data. These include the introduction of mini-supercomputers to provide for solution post-processing, preparation of input for new problems, and storage of the working solution files. Distributed files are being employed. Longer range plans for the NAS program contemplate making further increases in supercomputer speeds by a factor of four or more with each new supercomputer generation. To prepare for this, a longer range exploration of mass storage is in process.

Simulation, modeling and other analysis tools have been and continue to be used. Various workload models have been defined. In addition, performance monitoring and various measurements yield values on existing storage parameters and data.

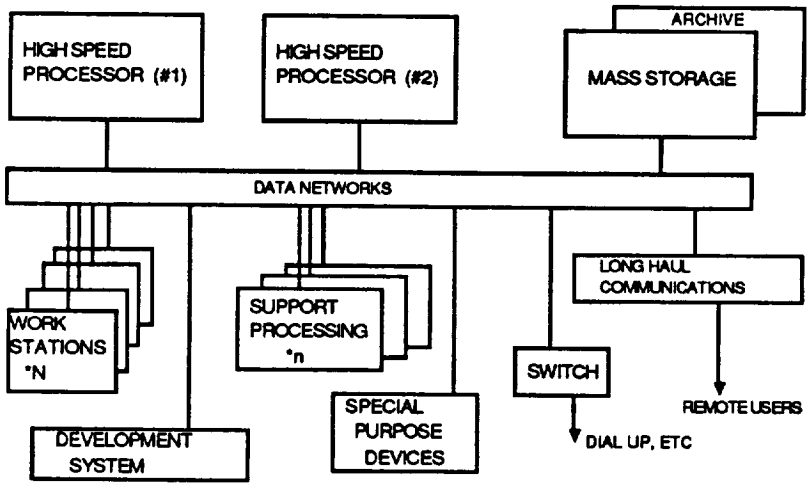


Figure 1 NAS SYSTEM CONFIGURATION (1988-89)

## WORKLOAD DESCRIPTION

The manner of using a system is critical in determining its requirements. Before discussing results of studies of requirements, an examination of the workload is desirable. Some of the methods of operation are somewhat unique to CFD problems and some are applicable to a broad class of problems. The storage requirements are similar in function to most other scientific applications, but may exceed many in volume of data produced which must be post-processed, retained for periods of months and brought

back from storage for searches.

The primary way to study solution results of these massive files is to generate three-dimensional, colored images and display them on a workstation. Displays may be a single image viewed from various aspects or a series of images generated as the scientist searches through results. To see the dynamics of a steady state situation, or for time dependent problems, an animated series of images is created, stored in various media, and played back in a simulated movie mode. A single illustration [2] of one image is shown in Figure 2.

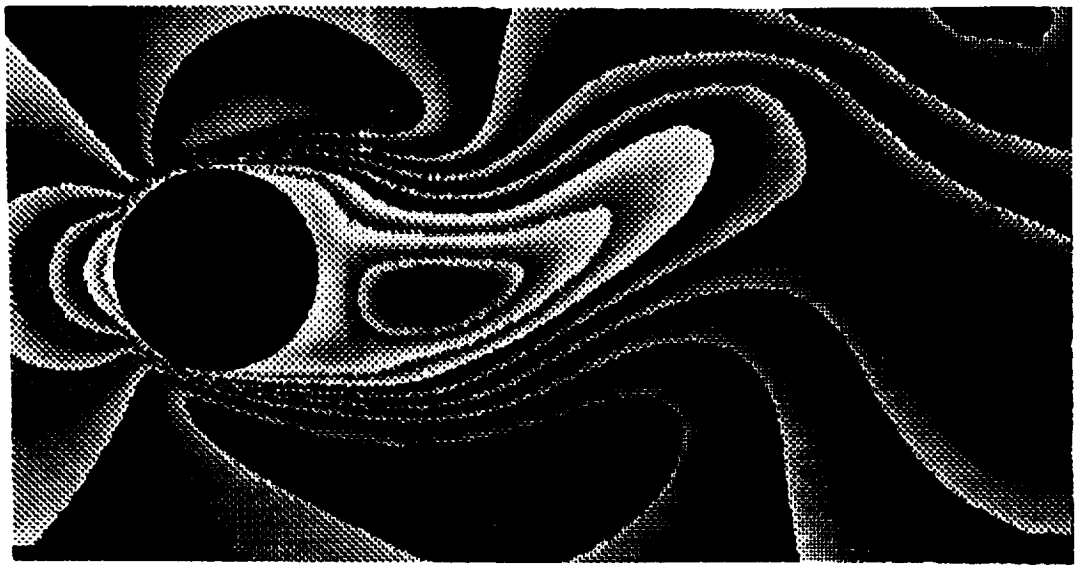

Figure 2 IMAGE OF SOLUTION SNAPSHOT

Before the NPSN was initially designed, there were a number of scientists at Ames working on CFD problems using various computers (Illiac, CDC-7600, Cray-1S). Extensive studies of their work and measurements produced an accurate and very detailed definition of their needs for that time period. [3 and 4] Computer memory sizes limited problem sizes so it was necessary to extrapolate the data to future supercomputers. Display and workstation capability also restricted those users and the methods of working was projected from the 1980 period to using workstations possible in the 1985 to 1990 era. These changes produced the workload basis for the Initial Operational Configuration of the NPSN.

For mathematical analysis and modeling, this workload data base gave a good projection of requirements. For simulation studies it was further expanded by breaking each user operation into detailed subtasks. These models served for initial studies but were not easy to adapt to other configurations or change for more advanced CFD research and

development. For expanding the NPSN, it was decided to use a much simpler workload model which included only those jobs that made the major demands on resources and neglected items that contributed less than about 10% to the total demands. The emphasis was shifted from workload to alternative configurations and modes of operation.

Scientific users, including the NAS users, are creative in the use of resources and adapt to available storage. They will move files around and delete material to keep within quotas. They may compromise by looking at a part of the problem, reduce a three dimensional data base to two-dimensional slices, or save only a fraction of their solutions. Measurements of existing usage can not be used alone in extrapolating future needs. These can best be assessed by talking directly with users and asking "What if . .. " questions. From these various sources, a projected workload and methods of operation have been developed.

Starting from the solution, the file requirements can be developed. The normal solution requires iteration to solve the non-linear differential equations that represent the physical relationship. To converge to the correct solution (or converge at all), the resolution must be fine grained in space (and time step if time dependent). The scientist experiments with a grid spacing until satisfied. The grid is usually non-linear with points more closely spaced in areas where high resolution is needed. The number of points imposes large demands on computer memory size. As a general case, from 50 to 100 points are needed per coordinate axis. For two dimensions this gives 2,500 to 10,000 points which are easy to accommodate. Adding the third axis increases these to 125,000 to 1,000,000 for the study situation of most interest. The number of computations increase much faster than linearly with grid size. This and memory now act to limit the scientist from using as many points as really needed.

The solution requires a minimum of eight values per point where three of these are spatial coordinates (x,y,z), and five are physical values (mass, three components of momentum, and energy). The coordinates are generally in the grid file and are stored once unless the problem requires adaptive grids. It may take 250 to 500 iterations to converge. This may mean hours on the supercomputer with the task broken up into a number of separate jobs. The last iteration is saved to restart for further iterations toward the solution. A user would most probably use 8 words per point for input and 5 for output. Words usually must be 48 bits or more; with 64 being the most used. For a million node problem, storage in words per solution would be 64 MegaBytes for input and 40 for results. This could

range downward in some cases but users would use more if available. More words per point are needed in the working memory during the solution. These requirements determine how many different problems can be in the supercomputer for solution and queued waiting to start.

Once the first solution has been examined, probably then modified and re-run until deemed proper, the scientist may change conditions to generate from a few to a large number of solutions over days or months. If the problem is static or repetitive flow, an animated movie or several might be produced using 100 to 200 steps for one repetitive cycle -- frequently saving the graphics file in addition to the solutions. A scientist will need many different looks at the solution data to examine different locations and parameters. As many as ten or more sets of graphics files will be produced for a single solution.

For time dependent problems, each solution after the first can be produced with only one iteration per time step. Time steps to yield time accurate results may be required at shorter intervals than needed to visualize the results. Thus, the user may wish to look at, say, every fifth iteration. For a one or two minute movie, from 1800 to 3600 steps would be needed for post-processing. These could be produced in about the same time as for two to six static problems. Five or more movies may easily be required over the months of study. A typical time dependent problem might take twenty times the computer time as a static situation and generate 1000 times the solution file quantity. The 40 Megabytes for a static job then could become 40 or more GigaBytes for a single time dependent analysis.

Solutions are post-processed to generate graphic images. Essentially, the physical variables of a solution are converted to graphic objects collected in display lists. The post-processing can be done in the supercomputer, in a mini-supercomputer, or in some workstations. To use limited resources better and speed up the interactivity, some schemes distribute the processing. Many workstations contain dedicated hardware and software for geometric transformations. The actual methods used depends on the problem and the equipment available. This processing generally reduces the storage needed by a factor from 5 to 40 for the graphics image compared with the solution. Images may be displayed to show pressures, temperatures, or a variety of physical values -- usually superimposed on the solid body moving through the fluid. It takes about 0.5% of the computations to generate one display list compared with the a single solution. The corresponding figure for time dependent runs is about 7%. These values are increased by the number of times a solution is viewed. Not all graphics files are saved because the display lists are specific to a

given parameter to be displayed and they can be recreated easily.

Up to this point, file requirements per single user for a day, or perhaps a week, have been calculated. Not everyone does the same work or produces the same output. The NAS system has over 600 active user accounts with perhaps 100 active on a given day. If only half of these produced just one time dependent analysis run per week, the total storage output would be 1 TeraByte per week and potentially 50 TeraBytes per year. This number is given just to demonstrate that total storage requirements involve gross estimates of job mix and other factors. How much is retained depends upon aging on the supercomputer, project duration and other factors best determined by experience. This is a function also of the configuration and will change. These estimates should be done carefully. Data expressed as average file sizes are meaningless. The majority of the files may be small, but the bulk of the storage is required for the large solution files. A number of files totaling 500 MB added to 20 GB does not change the storage requirements significantly. The mass storage designer would like to design for the near worst possible case but must compromise.

When the project analysis has been completed, the user will need to save the final results and discard files no longer significant. A major reduction in the amount of storage might result from this screening which only the file owner can do properly. These final results would probably be migrated from the mass storage to the cartridge archive.

Storage available to users will depend on the system configuration which in the case of the NPSN is changing with increases in supercomputer power. Experience has shown that the workstation user would like his own dedicated disk on the supercomputer as well as a processor when doing interactive analysis. This ties up an expensive and scarce resource. The NAS design policy is to keep the supercomputers fully occupied and not put on them work that could be suitably done elsewhere. This policy reflects in the design of supporting sub systems with performance that makes full use of the supercomputers' capability. There needs to be a reasonable balance between the supercomputer power and the workstation. To unload the supercomputers of work that can be done elsewhere and to provide more support for interactive analysis, mini-supercomputers are being considered for the network. Although the various hosts are all tied to common networks, each mini-supercomputer would serve a cluster of about 10 workstations. The aggregate disk storage on these could approach the total in the Mass Store. This would give users faster access and response and could help to cut down on the

total storage needed compared with a similar in crease in the Mass Storage.

## ANALYSIS MODEL

Various models and simulations were used for the initial design with rather limited utility -- largely caused by complexity. A designer would like something directly usable and easily changed. "Back of the envelop" calculations will not work because of the complex interdependence between various subsystems and method of operation. Jobs that can be queued and processed in the supercomputers are whole number values and not fractional. For the type of solutions used in the workload model, this number is relatively small to match the number of processors. This introduces abrupt changes in performance as the grid size is in creased. Further, for present algorithms, the number of calculations per grid point is now about the 3/2 power of the number of points [5]. See Figure 3. This led to the decision to make a simple model to aid designers and quantify a uniform set of conditions. In this paper, these values which are representative of a typical heavy use of the system are called the nominal baseline. The model user can vary from these as needed for some study. The basic nominal baseline values are given in the table in Appendix A.

A system model was designed to use a spreadsheet on a personal computer using Lotus-123 and then translated to EXCEL. Designers at IBM PC equivalents or Apple MAC's can work with the model, see all the assumptions, input values, and the resulting loading on processors, memories, disk storage, communication networks, and evaluate overall and subsystem performance. The model has about 375 equations and runs fast enough to give nearly instantaneous response. The designer can vary parameters and plot results. With some familiarity, the model user can change assumptions or equations. The model is modular so it can be changed to different configurations by someone who knows the model thoroughly.

The model mirrors the NAS operation which keeps the supercomputers fully productive. This determines the inputs for the other subsystems in terms of solutions per day and the amount of results to be analyzed. Because the model contains no queueing, it must be assumed that resources keep up with demands on a daily average basis. For results to be meaningful, the model user must adjust the support facilities to handle this output. The model results present statistics to allow the model user to make adjustments and try alternatives.

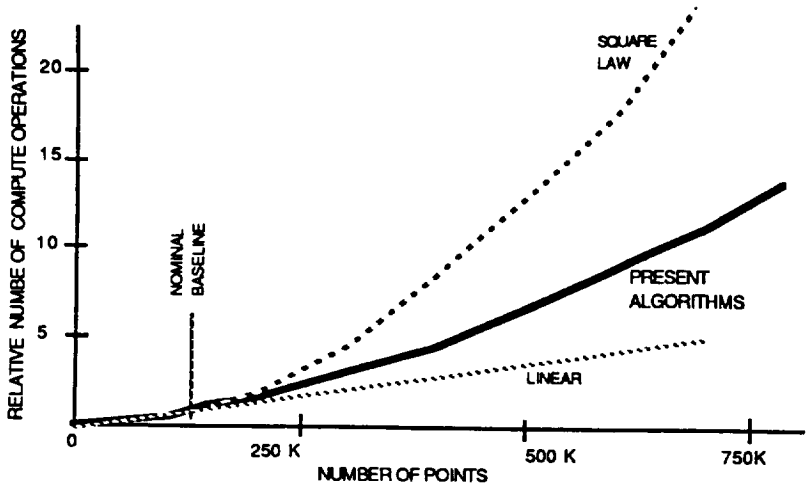


Figure 3 COMPUTATION GROWTH WITH GRID SIZE

The primary assumption employed in the model, in addition to full utilization of supercomputers, is to include major contribution to resource utilization and leave out minor tasks, files, etc. Provisions are included to allow for these by reserving some fraction of the resources. Workload and system assumptions are written into the model and displayed with a new line used for each. Input data values are likewise introduced one at a time. Each calculation depends only on values which have previously been defined or calculated. The model is organized by subsystem with a summary of results at the end of each section. A summary of system and more important subsystem values is at the end of the model.

Copies of the model are available to designers or analysts. It is described in a report that will aid the first-time user. After becoming familiar with the model, the documentation in the model should be sufficient for the user. The cell equations may of course be examined and these are supplemented by the equations written out in "english" plus extra notes.

## MASS STORAGE REQUIREMENTS

This paper does not attempt to develop the basis for any overall set of mass storage specifications. Some heavy-use situations will be shown with either a typical requirement or a parametric exhibit of how one value is influenced by input conditions and mode of operation. The values used are based on the nominal baseline unless otherwise specifically noted.

Mix of type of solutions A time dependent solution requires 10 to 20 times the processing time compared to a static solution. The time

dependent user needs 100 to 1,000 times as many solutions. Thus in a given time period the user working on some problem with values changing with time in some non-repetitive nature could create 10 to 50 times as many solutions. For these a large cost has been spent in computer time. The scientific analyst will need to save these solutions. Many months may be dedicated to working with the solutions.

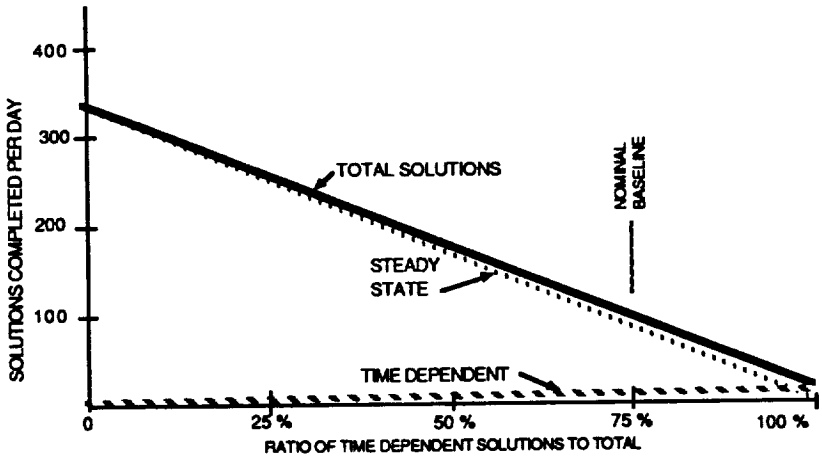


Figure 4 USE OF SUPERCOMPUTER PROCESSING POWER

In the baseline model, the parameters used show that 23 steady state solutions are possible for the same supercomputer resources as one time dependent set. This is illustrated in Figure 4. In the recent past, static solutions represent probably from 95 to 100% of the HSP usage. This will change as more processing power is available, with the forecast that up to 75% of the processing will be used for time dependent analysis. Therefore the baseline value for the model is set at 25% steady state and 75% time dependent. With one third as much CPU power devoted and the 23 to one ratio, the number of steady state studies supported would be about 7 times the number of time dependent. The baseline parameters result in about 86 steady state solutions and 11 time dependent runs com pleted in an average day.

The mix of studies by type performed is probably the major parameter in determining the solution files created, for a specific supercomputer. To examine the impact on this for input to the mass storage system, see Figure 5. The mass storage file input for the baseline conditions is 65 GigaBytes per day. The value in late 1987 with one supercomputer was in the range of 2 or 3, and this can be expected to be doubled in early 1988 and continue to grow. A twenty fold increase must be provided for or else users will have to make compromises.

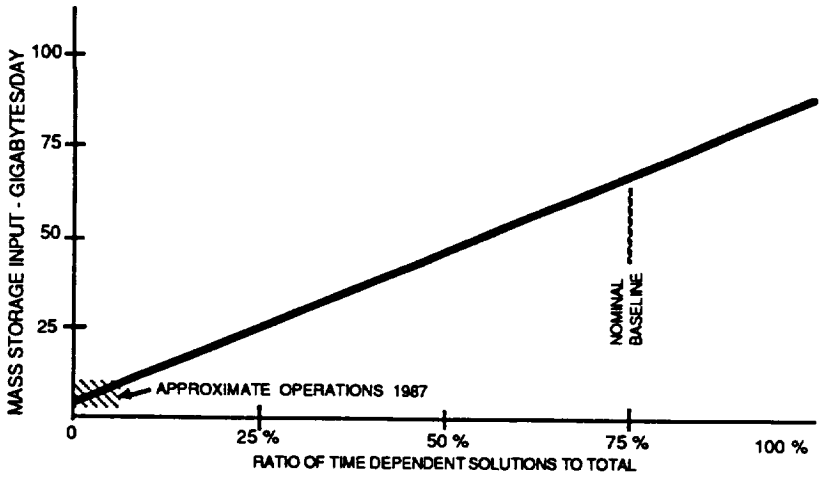


Figure 5 TOTAL SOLUTIONS PER DAY

Variation with Grid Size More complex situations demand better approximations to the physics and higher resolution in the analysis. This pushes the scientist to more grid points. In some situations it has been necessary to go to over 750,000 points to achieve a reasonable match to the "real world". Experience with today's algorithms were used to validate the model. The baseline uses 140,000 points. Figure 6 shows the variation of solution files sent to mass storage as the grid size is increased. Because a larger grid requires more time to converge, the number of solutions decreases faster than the grid points increase (See Figure 3) The amount of storage required decreases to about 27 Gigabytes per day for a million point problem or about one-half the value at 140,000. The resolution at 1 million points is only about twice the 140,000 case. Users wish they could work at well above a million.

The number of solutions per day goes down rapidly as the grid size is increased. For the mix of 75% time dependent jobs, the total is 97 at 140,000 nodes (nominal baseline). This drops to 5.1 at 1 million nodes -- the 1/20th value one would expect. (See Figure 3) This highlights a problem that has existed and continues to plague most supercomputer centers. Users would like to have more compute power, more memory, and more storage. If they were allowed this, then less solutions are produced. This results in supporting less users. The cost per user and per problem is increased rapidly as the solution time increases. This presents a problem for the management of the supercomputer center to balance the accounts and to obtain funding if only a very small number of users are served, even though these may be served very well.

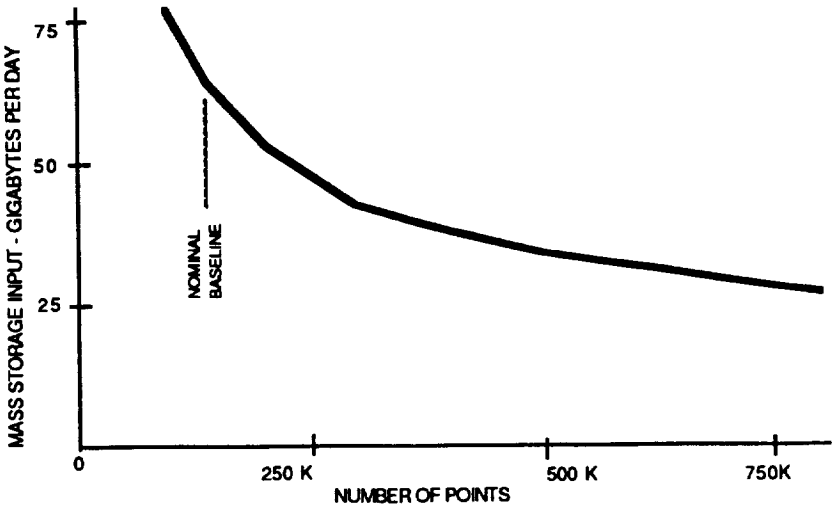


Figure 6 STORAGE REQUIRED VARIES WITH RESOLUTION

Other factors The model and the our analysis covers many other factors. Some of them have a significant impact on requirements for other resources such as network data transfer, temporary storage, support processing, and special purpose items, etc. For the mass storage, the factors discussed above cover the driving forces on the input rate. This paper can't go into the interesting but less significant contributors. With such wide variations possible the fine points are lost in the bigger picture.

Movement of the data is an important factor in the storage requirements. Many systems can not handle anywhere near 65 GigaBytes of new data each day. If it can not be delivered to the storage system input, it can't be saved; and, likewise the channel rates internal to the storage system must handle this. The traffic will contain peak rates much higher than the average. This boosts the required transmission rates plus perhaps imposes some need for a temporary buffer store. The file system must find and address the files. This is both global to the entire system and local to the mass store. Various hierarchical levels must be supported with access times that generally must be longer for the less frequently accessed data and archive files.

Sixty GigaBytes would overflow the storage on the NAS supercomputers in less than a day. It will fill the present mass storage in less than two days. The storage system must be able to migrate the files from one level to another. At 65 GigaBytes per day, solutions probably are best handled by sending them in a pipe-line mode on a dedicated channel directly to the storage system. With this mode, it is necessary to send copies of the solutions (or portions) to the support processors and workstations.

Related considerations This paper does not treat the data transmission considerations. However, files come in various sizes from small to medium to very large. The pipeline for time dependent solutions suggests separating that traffic. The remaining traffic still covers a wide range of sizes and related differences in latency time tolerated. These demands are not met well in a single network. It seems necessary to separate the large storage traffic from short messages/commands and interactive traffic to serve both well. Flow of data to the mass storage will probably come via various paths with differing characteristics and requirements.

Longer term aspects It is easy to project future supercomputer requirements for the next several generations. They need to be much larger and they will continue to be too slow and too small. Whatever they are, they will overload the storage and networks in speeds and total volume. Specialized computers will assume a bigger roll. The NAS charter is to keep abreast with the leading edge of these supercomputers. The design and execution of an increase by a factor of about five is now well underway. The next generation could bring a factor of four over that; or a total of twenty times the capability existing at the end of 1987.

This paper does not address beyond the factor of five step now in process. The analysis methods are probably not suited to any bigger step. Methods and technology now available do not appear to be suited to coping with storage inputs exceeding 250 GigaBytes per day.

## OBSERVATIONS

With increasing supercomputer speed, more solutions are produced. With this comes the need to proportionally increase the mass storage capability. As supercomputing power is added, there is a dilemma in allocation of this to projects. The common mode is to allow a large number of users accounts and this forces smaller than optimum grid sizes or very long turn-around times. It also increases the amount of storage files created by a significant factor. Users have responded by not doing very much time dependent work and not saving all the solutions.

When a new, faster supercomputer becomes available the dilemma is most apparent. Probably better science per project but less projects would be done if bigger faster supercomputers were limited to only jobs too big to fit on existing machines. The actual allocation becomes a trade-off with the funding politics playing a role.

The mass storage requirements faced within the next two years exceed space and cost limits. The designers and users must compromise and develop alternative methods of working in a constrained resource environment.

Today's technology is not up to today's demands. The addition of mini-supercomputers and very high performance workstations will allow the high speed processors to be devoted mainly to generating solutions which can only be done there. With these mini-supercomputers, there will be added working storage separated from the supercomputers and from the mass storage system. Mass storage would then be largely for permanent storage at various levels with different media and speed of access.

New technology can be expected to increase the storage densities available. The major developments in this are driven by the marketplace and this has not focused on supercomputer center requirements. New methods of working with the data and the presently known storage density limits can achieve factors of two or four in the immediate future. More is required as soon as possible.

Improved algorithms are continually being developed. Advances in these over the past 15 years rivals the increased performance achieved in computer hardware speeds and cost reductions. This area of progress is likely to reduce the computational time and thereby actually increase the storage requirements. Progress in reducing the data required per solution is not apparent and may be very difficult. Some savings are possible with data compression but floating-point numbers are notably hard to compress.

Some day, it may become possible to raise the level of the CFD from solutions done in hours to near real-time. This will permit going from the present largely batch produced solutions to a simulation mode. When this is done, the solutions might not be saved. Instead the steps to re-create the solution plus the graphics would be saved. The latter lend themselves to considerable compression. At this unknown time, the storage requirements will be different and may be less of a problem.

## CONCLUSIONS

Within the next two years, the NAS system faces the need to be able to handle inputs to the mass storage in the order of 50 GigaBytes per day. The total accumulation for a year would then be around 15 TeraBytes.

Not all of this storage can fit within disk drives because of physical space and economic constraints. A robust archiving system will be necessary to augment the disks.

New and alternate procedures will be needed to provide good service to users. This includes moving time dependent solutions during the computation of the next time step.

Improved technology is needed to meet the challenge of the huge volume of data storage needed for scientific applications.

## ACKNOWLEDGMENT

This work benefits from significant contributions made by Dr. Thomas Lasinski who is Workstation Subsystem Manager on the NAS program. He contributed to the definition of the workload from the scientist's view and worked with the model to help validate it.

## REFERENCES

[1] B. Blaylock, et al Extended Operating Configuration Design and Development Plan NASA Ames (June 1987)

[2] C. Levit and D. Tristram of NAS project developed program cplane used in Figure 2.

[3] D. Chapman, H. Mark, and W. Pirtle, 1975 "Computers vs. Wind Tunnels." Astronautics and Aeronautics (Apr. 22-35)

[4] R. Levine, "Supercomputers", Scientific American Jan. 1982 (pg. 118)

[5] W. Van Dalsem, NASA Ames Internal memos: Power law analysis.

Appendix Table follows

## APPENDIX A TABLE OF MODEL NOMINAL BASELINE VALUES

(PARTIAL LIST)

WORKLOAD VALUES (Parameters subject to model user change)

GRID SIZE = 140,000 POINTS RATION OF HSP-1/HSP-2 = 1.0

EXPONENT OF COMPUTATION GROWTH WITH NUMBER OF POINTS = 1.5

AVERAGE NUMBER OF ANALYSIS VIEWS PER SOLUTION = 10.0

RATIO OF TIME DEPENDENT HSP TIME TO TOTAL TIME = 75%

RATIO OF SOLUTION TIME TO TOTAL HSP TIME = 91%

RATIO OF SOLUTION TIME ALLOWED FOR SUPPORT PROCESSING = 10%
(Balance of support processing done on mini-supercomputers)

RATIO OF USE OF HSP TIME FOR REMOTE USERS TO TOTAL = 55%

FRACTION OF SOLUTIONS SAVED TO MASS STORE = 1.00

FRACTION OF SOLUTIONS RETRIEVED FROM MASS STORE = 0.1

FRACTION OF SOLUTIONS SAVED AT END OF PROJECT = 0.75

SYSTEM OPERATION HOURS = 24 HSP's = 22 (Secure mode = 3)

SYSTEM VALUES (Parameters match design values)

HSP-1 = 250 MFLOPS WITH 4 PROCESSORS AND 256 MILLION WORDS

HSP-2 = 1,000 MFLOPS WITH 8 PROCESSORS AND 256 MILLION WORDS
(For above, multiprocessing = YES )

HOURS PER DAY INTERACTIVE USE OF SUPPORT MINI-SUPERCOMPUTERS = 12

HOURS PER DAY INTERACTIVE USE OF WORKSTATIONS = 12